

Språkbanken och Korp: Mot en språkteknologibaserad forskningsinfrastruktur

Lars Borin
Språkbanken/svenska språket, Göteborgs universitet
Giellatekno, UiT

19/2 2014

Språkbanken: lite historia

- ~1970: första svenska korpusen: **Press-65**
- 1972: professor i **språkvetenskaplig databehandling**
- 1975: **Språkbanken** ("Logoteket")
- 1984: **datalogvistikprogrammet**
- 2000: **GSLT** (forskarskola i språkteknologi)
- 2004: pilotprojektet **Litteraturbanken**
- 2007: **CLT** (Centre for Language Technology)
- 2008: **språkteknologi** styrkeområde vid GU
- 2009: **strategiska GU-medel** till styrkeområdet **språkteknologi**
- 2011: svensk partner i **META-NORD**
- 2013: nationell samordnare för **SWE-CLARIN**

Språkbanken – vad, för vem, till vad?

vad är Språkbanken?

- ▶ en nationell resurs sedan 1975
- ▶ en FoU-enhet i språkteknologi (med nationella och internationella samarbeten, t.ex. EU-projekten **META-NORD** och **CLARIN**)
- ▶ **fri tillgång** till sökning i digitala, förädlade språkresurser (svenskt skriftspråk av alla genrer från alla tider):
 - ▶ textkorpusar (enspråkiga och parallella)
 - ▶ lexikonresurser (moderna och historiska)
- ▶ unik kompetens inom området svenska språkresurser

(traditionellt) för vem och till vad?

- ▶ språkforskare (nordister och lingvister)
- ▶ lexikografer
- ▶ språkteknologiforskare
- ▶ utbildning
- ▶ allmänheten

Språkbanken: <http://spraakbanken.gu.se>



The screenshot shows the Språkbanken website interface. At the top, there's a navigation bar with links like 'Om oss', 'Resurser', 'Forskning', 'Publikationer', 'PhD program', and 'Personal'. Below this, the main content area is divided into several sections:

- Språkbanken**: A brief introduction to the institution and its mission.
- Nya ord i SALDO**: A section highlighting new words in the SALDO lexicon, with a date of 2014-02-03 19:32.
- Forskning**: A section for research projects, including 'SweFNN++', 'Kulturomik', 'SweCan', and 'Diabasa'.
- Resurser**: A section for resources, including 'Korp' (Corpus) and 'Användarhandledning' (User manual).

On the right side, there are two tables showing statistics:

Korp	
antal korpusar	160
token (total)	1 711 278 707
antal meningar	119 449 334

Korp	
antal lexikon	22
ingångar (totalt)	692 727

Ansikten utåt 1: Korp

Moderna | Parallella | Fornsvenska | Litteraturbanken | Spf 1800-1900 | Aldre finlandssvenska | Färöiska | Digidaly Svenska | English

KORP

346 korpusar valda — 1 808 399 306 token

Enkel Utökad Avancerad

Sök efter: Sök även som ☐ förled ☐ efterled och ☐ skiftelagesberoende

Relaterade ord:

förlova tappa_bort förlova förspilla ge_tillspillo mista tillspillo borttappa till_spillo oförlova släva_bort tillspillo ge_till_spillo

lida_nederlag förlova förlova förlova förlova ge_app oförlova bita_L_graset ge_app uppge

KWC: träffar per sida: 25 sorterat inom korpus på förekomst Statistik: sammanställ på: ord

KWC Statistik Ordbild

Antal träffar: 117 476

Förslagna: 1 2 3 4 5 6 7 8 9 10 11 ... 4899 4700 Nästa Visa kontext

Åbo Universitet sen 2012

De större partierna har att bokföra den som en oundviklig **oförlost**

Efter den knappa förlusten mot tjeckerna kan Finland

Korpus

Åbo Underrättelser 2012

testattribut

datum: 2012-09-28

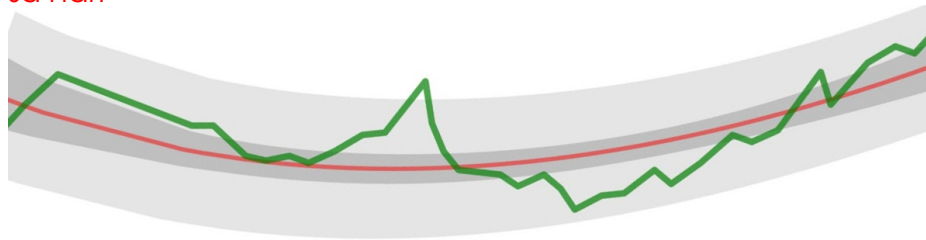
ordattribut

Korp-ideologin

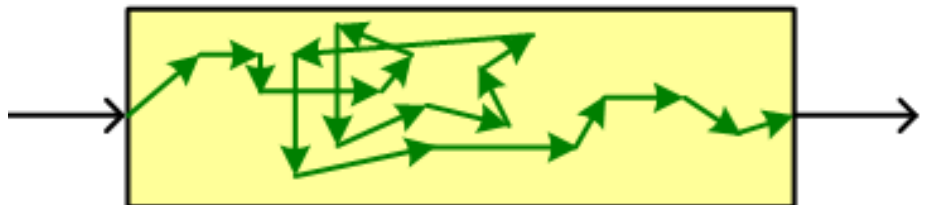
- ▶ Korp är Språkbankens korpusinfrastruktur
- ▶ Korp har en **teknisk sida** och en **användarsida**
- ▶ De tekniska lösningarna ska vara bra för användarna i stort och på lång sikt,
- ▶ vilket innebär en balansgång
 - ▶ Att bygga solida tekniska lösningar som är generella tar ibland lång tid
 - ▶ medan en sårlosning för ett individuellt fall kan åstadkommas relativt snabbt

rörelsen framåt är viktig

så här:



men inte så här:



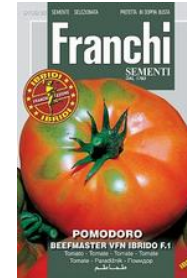
tekniska lösningar

- ▶ Den viktigaste tekniska lösningen i Korp och dess syskon är
 - ▶ att korpussökmaskineriet är **strikt separerat** från de program som använder det, **inklusive själva sökgränssnittet**
- ▶ Vi talar om
 - ▶ Korps **bakända** och
 - ▶ dess **framända**
- ▶ Det betyder att man kan ha ett godtyckligt antal gränssnitt för olika grupper och olika behov och
- ▶ "användaren" är typiskt inte en människa, utan ett datorprogram

tekniska lösningar, 2

- ▶ Nästa viktiga tekniska lösning har att göra med "ingångarna" till bakändan
- ▶ Det gäller att hitta rätt frihetsgrad/abstraktionsnivå
- ▶ för då kan man blanda och ge på ett väldigt produktivt sätt
- ▶ Kanske man bäst tänker på bakändan som en samling funktioner som man kommer åt genom ett standardiserat gränssnitt.

abstraktionsnivå/frihetsgrad



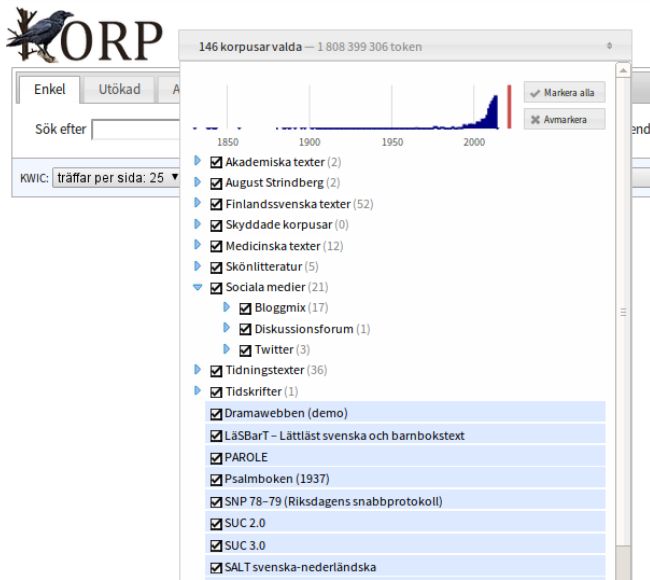
tekniska lösningar, 3

- ▶ När man börjat tänka så
- ▶ blir det naturligt att göra så många funktioner som möjligt tillgängliga på samma sätt
- ▶ inte bara söksystemet, utan även korpusimport- och -anoteringsfunktionerna
- ▶ Man tänker förhoppningsvis mer i termer av modularisering och återanvändning
- ▶ **MEN** detta arbetssätt kräver en mycket hög och solid teknisk kompetens

användarfunktioner

- ▶ Detta finns nu i Korp:
 - ▶ KWIC-visning
 - ▶ tidsuppmärkning och funktioner för att använda den
 - ▶ annotationer: ordklass/msd, lemgram, dependenssyntax(, ordbetydelse)
 - ▶ statistikfunktioner (tabell, tårtdiagram, trenddiagram)
 - ▶ ordbild
 - ▶ bortåt två miljarder ord moderna texter, och nästan en miljard ord äldre textmaterial (i Korplabbet)
 - ▶ nedladdningsbara "meningsmängder" (slumpvis omkastade texter)
 - ▶ möjlighet att lösenordsskydda korpusar och funktioner för användaradministration
 - ▶ all mjukvara (bakända och framända) fri och nedladdningsbar för egen installation

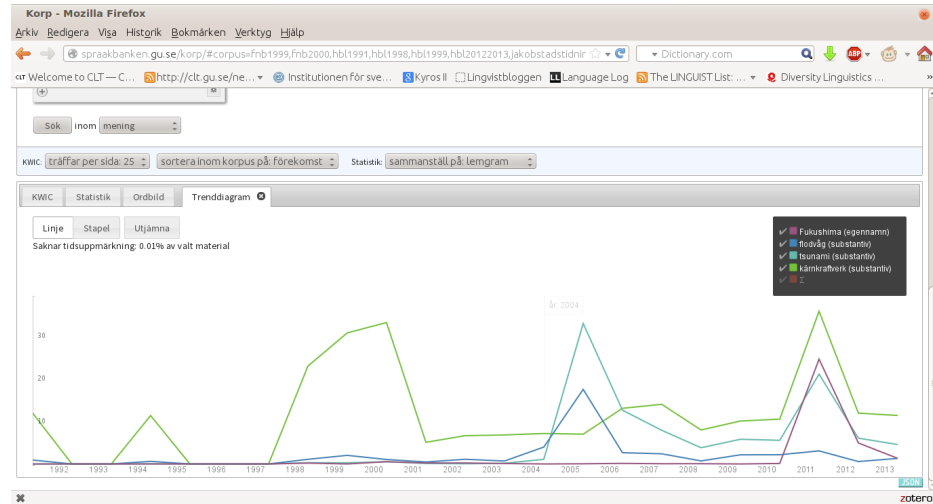
Korp: korpusval



Korp: KWIC-visning

[illegible]

Korp: Trenddiagramm



Korp: Ordbild – surfa (verb)


Subjekt	surfa	Objekt	Adverbial		
1. du	40	1. —	40598	1. på nät	250
2. treåring	24	2. porr	55	2. på internet	97
3. anställda	30	3. nät	32	3. bland blogg	45
4. svensk	33	4. våg	23	4. lite	117
5. kund	25	5. 3g	18	5. på sida	73
6. proc ²	21	6. stund	29	6. på hemsida	65
7. hälft	15	7. timme	37	7. utomlands	47
8. användare	15	8. nätbutiker	8	8. på våg	25
9. besökare	14	9. sida	33	9. på facebook-konto	16
10. folk	31	10. hemsida	23	10. i shop	20
11. procent	22	11. internet	20	11. på hemnet	20
12. emanuelkarlsten	12	12. datavolym	6	12. på Internet	26
13. mp3-bok	5	13. psl	6	13. i cyberspace	20
14. man	26	14. skånelängor	6	14. på blogg	43
15. våg	9	15. 9gag	6	15. stund	48

Korp: Ordbild – förlust (subst.)

Preposition	Attribut	förlust	Efterställt Attribut	Förlust	verb	Verb	förlust		
1. efter	5699	1. rak	4057	1. på krona	5079	1. innebära	717	1. redovisa	2808
2. med	7544	2. stor	5345	2. på dollar	1060	2. bli	1442	2. göra	3076
3. trots	1363	3. tung	1324	3. före skatt ²	908	3. uppgå	505	3. lida	767
4. utan	1641	4. ekonomisk ²	1178	4. före skatt	908	4. vara	4140	4. täcka	601
5. mot	797	5. ekonomisk	1178	5. efter finansnetto	658	5. svida	194	5. innebära	715
6. i och med	98	6. snöplig	289	6. på miljon	714	6. komma	1041	6. orsaka	389
7. dagen efter	75	7. knapp	403	7. på euro	628	7. betyda	221	7. vända	391
8. över	405	8. hedersam	199	8. på match	505	8. redovisa	159	8. vända ²	391
9. vid	702	9. svår	503	9. för period	309	9. var än	171	9. inkassera	196
10. på grund av	116	10. bitter	247	10. på miljard	323	10. beräkna	142	10. skylla	232
11. även om	34	11. enorm	272	11. för kvartal	335	11. bero	160	11. tillfoga	126
12. förutom	68	12. klar	255	12. av människoliv	129	12. landa	91	12. ta	765
13. nära	66	13. smärtsam	132	13. av mångfald	147	13. landa ²	91	13. medföra	176
14. till följd	24	14. eventuell	220	14. på mark ²	204	14. göra	409	14. undvika	205
15. efter	6	15. oväntad	155	15. av arbetstillfälle	110	15. kännas	158	15. visa	348
		16. raka	10						
		17. raka	10						
		18. rak	10						
		19. sovjetisk	3						
		20. snöplig	1						

Ansikten utåt 2: Korplabbet

Moderna | Parallella | Förrsta | Litteraturbanken | Språk-2000 | Aldre frändska | Klopptiska | Rimebiter | Bibelstallet | Lagrummet | Digitala | Historiskt

ORP  5 korpusar valda (alla) — 23 538 445 token

Enkel Utökad Avancerad

Sök efter Sök även som ☐ förfed ☐ efterled och ☐ skiftlagseoberoende

Relaterade ord

domsrätt etradsrätt svärst rättfärdig inkränkning anrätt rätteligen krögrätt förfoganderätt dispositjonsrätt älskrift
förhandlingsrätt rättslig bevisrätt rättslös församlingsfrihet beutingsrätt självskatt rättsva statsrättslig processrätt uttänderätt
lösningrätt stalsrätt familjerätt nödrätt älskriftsrätt äganderätt extentionalrätt rättfärdiga

Visa fler

KWC: träffar per sida: 25 sorterar inom korpus på: förekomst Statistik sammanställ på: ord

KWIC Statistik

Antal träffar: 7 336

Förhandsvisning: 1 2 3 4 5 6 7 8 9 10 11 ... 293 294 Nästa Visa kontext

Dröms röttörre

Iedd närmast af skäligen emkä taktiska hänsyn, **proklamerade** lösen: allmän **öfver** ingen » nationell samling » säga de rösträttslösa likt självförlidiga pojkar, som vilja mutas för att vara snälla.
Utan allmän **rösträtt** och samling i det svenska » franses hus » under rösträttsdebatten anno 1906 — och man hoppas, att det
Hviken är då för dem den reella betydelsen af allmän **rösträtt** — om man försöker se fullkomligt nyktert på saken?
Men hvilkens glädje får då industriarbetaren af den allmänna **rösträtten**?

Korpus
Diverse tidningar
textattribut
titel: Det Nya Svealand
datum: 1907
ordattribut
grundform:
texttaggar

Ansikten utåt 3: Korps annoteringslabb

ORP Annoteringslabbet

Ladda exemplet: [Branäs](#) [Ålåsabo](#) [Täbörken](#) [Lindarö](#) [Exempelortspolis](#) **Sprik:** [Svenska](#) [Engelska](#)

1 Att en fågel så kan svara, hör ju till det underbara!

Visa avancerade inställningar

VB: Verb

ord	msd	lemma	lex	saldo	prefix	suffix	ref	deph	deprel
Att	SN	att	att.sn.1	att.2			01	08	SS
en	DT. UTR. SIN. IND	en	en.al.1	den..1.en.2			02	03	DT
fågel	NN. UTR. SIN. IND. NOM	fågel	fågel.nn.1	fågel.1	få..av.1. få..vb.1	gel.nn.1	03	05	SS

Ansikten utåt 4: Karp

KARP

23 lexikonresurser valda (alla) — 693 410 ingångar

Sök historik

Enkel Utokad Editor

hund (substantiv)

Sök via diapijot

Trafär per sida: 25

"Förleden Wecka är en liten svarstraggar **Hund** för agaren bortkommen, eho den upprigt, behagade tistalla densamma uti Madama Florins hus pa Konungsgatan."

Fullständiga träffar (41) Statistik

Filtrera bände: 1 2 Nästa

Saldo Sa... Se... Sv... Sw... Pa... Ke... Uv... Wo... Lexin Delns ordbok

Skrytja Söberna Söderma Di...

Saldo (2)

Betydelse	Lemang	Ordklass	Påminr	Sekundär	Barn (primära)	Barn (sekundära)
hund	hund (substantiv)	substantiv	djur	salislap	<ul style="list-style-type: none"> Karo - Pompe Karl XII bandhund benda 	<ul style="list-style-type: none"> andeogon panna apportera hämta bussf ämbala
	(230490)				Visa alla (33)	Visa alla (44)

hund^a hund (substantiv) substantiv using

(230490)

Ansikten utåt 5: Lärka

Övningsgenerator | Hit-Ex (meningsläsbarhet) | Redigeringsverktyg för inlämningspusslar | Svenska



Studenter i språkvetenskap | Öva satsdelar | 7 av 7 relationer valda

Helt automatisk

Resultatanslutar

Övningsgrupp: Korpus/totalt
Lingvistik/SVHT1, självstudier: 3/4

Öva satsdelar

Valj en korrekt satsdel till den markerade frasen

Num	Menning	Ditt svar	Rätt svar	Lärkor
4	Vattnet tas in och släpps ut i Öregrundsgrepen .	Valj relation		
3	Barnbidraget betalas fr.o.m. kvartalet efter det då barnet föddes .	adverbial	✓ adverbial	
2	Färre barn skaffar man sig först då man ser en utgå ur sin fattigdom .	indirekt objekt	✓ indirekt objekt	

pågående projekt: SweFN++



SweFN++

Sök i SweFN++

Publikationer

Utvecklingsversion

Dokumentation

Historik

Felrapport

FM-SBLEX

FrameNet Workshop 2013

NoDaLiDa 2013 workshop

Flerordsworkshop 19/3 2013

Lars Borin, Dana Dannélls, Markus Forsberg, Karin Friberg Heppin, Richard Johansson, Dimitrios Kokkinalakis, Leif-Jöran Olsson, Maria Toporowska Gronostaj, Jonatan Uppström, Kaarlo Voionmaa.

Följ utvecklingen via RSS.

Svenskt frasnät++ (SweFN++)

Detta projekt finansieras av VR/RFI 2011-2013 (dnr 2010-6013) samt med särskilda medel från Göteborgs universitet till styrk området språkteknologi (2009-2015).

SweFN++-projektet handlar om att skapa en central infrastrukturkomponent för svensk språkteknologi, nämligen en stor fritt tillgänglig lexikonresurs med rik lingvistisk information. Man kan säga att den planerade resursen kommer slå en bro mellan det förflutna och framtiden:

Det förflutna, därför att vi vill återanvända en rad fria lexikonresurser som har tagits fram i olika projekt vid olika tidpunkter av olika forskargrupper, men som sen har fått mindre användning än de förtjänar främst på grund av idiosynkratiska format och brist på driftsmedel för att underhålla resurserna;

framtiden, därför att vi till de integrerade befintliga resurserna vill lägga den typ av avancerad och mycket användbar semantisk och syntaktisk information om orden som man finner i det engelska Berkeley FrameNet (BFN) och några få liknande resurser för andra språk, ett arbete som vi planerar att göra i samarbete med den forskargrupp som står bakom BFN.

Eftersom dessa befintliga lexikonresurser representerar stora insatser i möda och pengar och eftersom de i många fall innehåller högvärdig språklig information, vill vi alltså rädda så mycket som möjligt av dem från förgängelsen samt vidareutveckla dem.

Finansierat av VR/RFI

SweFN++: SweFN

Cure mod

domän	Med
kärnelement	Affliction Body_part Healer Medication Patient Treatment
periferielement	Degree Duration Manner Motivation Place Purpose Time
exempel	<ul style="list-style-type: none">[Salvan]Medication [läker]Lu [[skrubbsår]Affliction och [brännsår]Affliction]Affliction .[Läkaren]Healer [botade]Lu både [[ryggskott]Affliction och [vatten i knät]Affliction]Affliction , innan hon hamnade i svårigheter efter att ha utfört ett förnyringsexperiment med patienten.[Genterapi]Treatment [botade]Lu [dödssjuka i cancer]Patient .[Läkaren]Healer [botar]Lu [kroppen]Body_part och [filosofen]Healer [botar]Lu [själen]Body_part , men det krävs ett engagemang för att lyckas.Traditionellt har [stafylokockinfektioner]Affliction enkelt [botats]Lu [med antibiotika]Medication .Syftet med terapi för en personlighetsstörning är inte att [fullständigt]Degree [bota]Lu [patienten]Patient , eftersom det varken är möjligt eller eftersträvsvärt.Han ansåg även att [Gud]Healer hade [helat]Lu [honom]Patient [från cancer]Affliction .[Transplantation]Treatment kan ha [botat]Lu [[hiv-smittad]Patient]Affliction .
lus	vb <i>botat¹ helat¹ läkat²</i> nn <i>läkning¹ botande¹ helande¹</i> av <i>botigt¹</i>
kommentar	My ram jämfört med BFN. Den ursprungliga tolkningen av ramen Cure i BFN ges här en snävare tolkning som implicerar att ett positivt resultat av någon form av medicinsk behandling föreligger.
skapad av	MTG
skapad	2012-04-02
ändrad	2013-12-09

digital areallingvistik

Digital areal linguistics

Word lists

The languages are shown with their names and ISO 639-3 codes in parentheses. In case a language has no ISO 639-3 code, nothing is displayed.

Languages

- Hindi (ISO 639-3 code: hin)
LWT: [\[link\]](#)
ISO: [\[link\]](#)
- Marathi (ISO 639-3 code: mar)
LWT: [\[link\]](#)
ISO: [\[link\]](#)
- Koraghi (no ISO 639-3 code)
LWT: [\[link\]](#)
ISO: [\[link\]](#)
- Telegu (ISO 639-3 code: tel)
LWT: [\[link\]](#)
ISO: [\[link\]](#)
- Bengali (ISO 639-3 code: ben)
LWT: [\[link\]](#)
ISO: [\[link\]](#)
- Punjabi (ISO 639-3 code: pan)
LWT: [\[link\]](#)
ISO: [\[link\]](#)
- Khasi (ISO 639-3 code: kha)
LWT: [\[link\]](#)
ISO: [\[link\]](#)
- Tamili (ISO 639-3 code: tam)
LWT: [\[link\]](#)
ISO: [\[link\]](#)
- Nepali (ISO 639-3 code: nep)
LWT: [\[link\]](#)
ISO: [\[link\]](#)
- Gujarati (ISO 639-3 code: guj)
LWT: [\[link\]](#)
ISO: [\[link\]](#)
- Kannada (ISO 639-3 code: kan)
LWT: [\[link\]](#)
ISO: [\[link\]](#)
- Sanskrit (ISO 639-3 code: san)
LWT: [\[link\]](#)
ISO: [\[link\]](#)
- Tibetan (ISO 639-3 code: bod)
LWT: [\[link\]](#)
ISO: [\[link\]](#)
- Kurdi (ISO 639-3 code: kur)
LWT: [\[link\]](#)
ISO: [\[link\]](#)

A (revised) Swedish ISO 639-3 code (swe) list is available for download in LWF format [\[link\]](#). It can also be searched online through [\[link\]](#).

License

This work is licensed under a [Creative Commons Attribution 3.0 Unported License](#).

Finansierat av VR. Ett samarbete mellan Språkbanken/Göteborg, lingvistik/Uppsala och lingvistik/MPI-EVA, Leipzig

Culturomics

Culturomics: core NLP technologies

Culturomics: language over time

Culturomics: publications

Culturomics: question answering

Culturomics: text processing in historical texts

Culturomics: text processing in social media

Culturomics: tracking semantic change

Culturomics: visualization

Culturomics: meetings

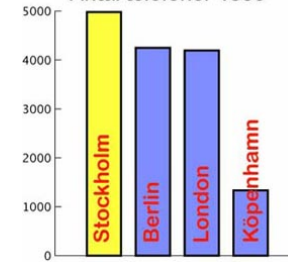
Exploring language over time

The following figures show the kind of results that emerge directly from a linguistically annotated text material available through Språkbanken's general corpus infrastructure. Unlike the culturomics work referred to earlier, the diagrams show the distribution of the lexemes (lexicon words) tsunami and flodvåg in a newspaper material covering the years 2001–2011, including all inflectional forms and all compounds containing these words. This is made possible by the lexical analysis tools based on handcrafted resources used for annotating Språkbanken's corpora.

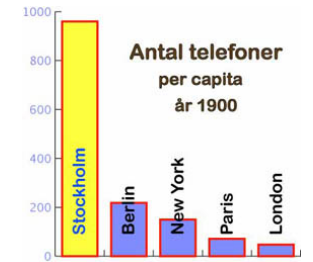
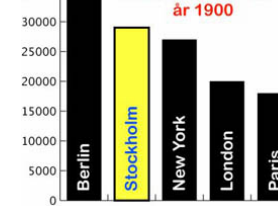


Finansierat inom VR:s ramprogram Det digitaliserade samhället – igår, idag, imorgon. Ett samarbete mellan Språkbanken/Göteborg, datavetenskap/Chalmers och datavetenskap/Lund.

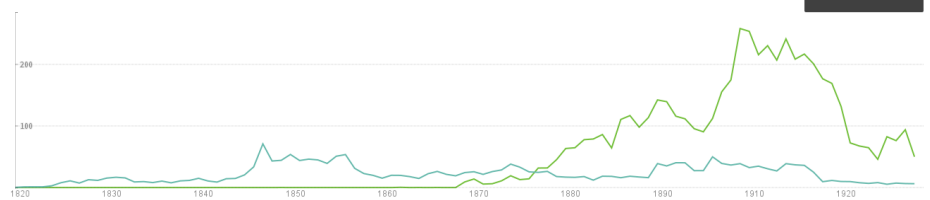
Antal telefoner 1885



Antal telefoner år 1900



Linje Stapel Utjämna



nya projekt: Koala

Infrastruktur

- Korp
- Karp
- Lärka
- SBLEX
- FM-SBLEX
- Koala

Koala – Korps lingvistiska annotationer

Projektet Koala -- Korps lingvistiska annotationer -- handlar om att utveckla en infrastruktur för text-baserad forskning med högkvalitativa annotationer.

Korpusinfrastrukturen Korp på Språkbanken (<http://spraakbanken.gu.se>) innehåller stora mängder text av olika typ och ålder, som används av forskare inom olika områden och av allmänheten. Texterna innehåller lingvistisk uppmärkning, annotationer, som ordklasser och syntaktiska roller, vilka hjälper till att filtrera sökresultaten för användaren. De låter oss hitta "sjöng" och "sjungit" när vi söker efter "sjunga" och alla ställen där Caesar är objekt till verbet besegra utan att vi behöver titta på dem där han är subjektet, samt att vi inte behöver titta på meningar om lokaler när vi letar efter "lounge", utan kan fokusera på förekomsterna som handlar om djuret. Annoteringarnas kvalitet är avgörande för att få bra sökresultat, särskilt för forskare som annars kan behöva gå igenom tusentals irrelevanta meningar.

Målet för Koala-projektet är att förbättra annoteringarna, som har skapats automatiskt med välkända språkteknologiska metoder. Det görs genom att lägga till språklig kunskap i systemet via de många resurser som finns tillgängliga via Språkbanken, samt genom att kombinera de olika annoteringsverktygen för lexikal analys, ordklasstaggning, betydelsedisambiguering och syntaktisk analys till ett högkvalitativt system där annoteringar på ord- och meningsnivå informerar varandra och där systemet inte fattar beslut innan det har all tillgänglig information. De data och verktyg som blir resultatet kommer att göras fritt tillgängliga.

Projektet finansieras 2014-2016 av Riksbankens jubileumsfond.

Finansierat av RJ/infrastruktur

nya projekt: MAPIR

Hem > Forskning > MAPIR

Webbkarta

Forskning

- Infrastruktur
- SweFN++
- META-NORD
- KELLY
- Kulturomik
- CONPLISIT
- Digital areallingsvistik
- ITG
- MOLTO
- PINCORE
- A System Architecture for ICALL
- Akademiska ordlistor
- Corpus-driven induction of linguistic knowledge

MAPIR

Svenska språket under medeltiden, fornsvenska (ca 1225-1526), finns bevarat i manuskript, brev och tidigt tryck. Dessa dokument är värdefulla för många olika forskare, såsom lingvister intresserade av svenska språkets förändring under tiden, juridikforskare som vill undersöka medeltida lagar, teologer som studerar tidiga översättningar av bibeltexter, eller medicin-historiker som är intresserade av medeltida folkläkeläkonst.

I MAPIR-projektet -- Metoder för automatisk Analys av Text i digitala Historiska Resurser -- skapar vi verktyg för automatisk lingvistisk analys av fornsvenska. Projektet är relaterat till Språkbankens satsning på historiska resurser, Diabase, och ligger inom forskningsområdet datalingvistik, vetenskapen om datamaskinell språkbehandling och datorstödd språkforskning. Genom att lägga till grammatisk information i digitaliserade fornsvenska texter underlättar vi studier av detta kulturarv och möjliggör nya sätt att undersöka det.

Att utveckla verktyg för fornsvenska är en utmanande forskningsuppgift, även med de främsta datalingvistiska metoderna. Detta beror på egenskaper i de fornsvenska texterna. För det första förändrades språket under den fornsvenska tiden vad gäller till exempel ordföljd och ordböjning. För det andra fanns ingen rättstavning i dagens bemärkelse. Samma ord kunde stavas på flera olika sätt. Ordet "mapir", som betyder man eller människa, stavades till exempel även "mæpr", "mander" eller "meber". Man kan till och med se olika stavningar för samma ord i ett enda stycke. För det tredje skiljer sig språket mycket åt mellan texterna. Det har gått 300 år mellan de äldsta och de yngsta texterna, och de kommer från olika geografiska områden och är av olika typ. För det fjärde kräver de flesta automatiska metoder antingen en mycket detaljerad datamaskinell beskrivning av ett språk, eller en större mängd text som redan har lingvistisk uppmärkning som datorn kan lära sig av. Inget av detta finns i dagens läge för fornsvenska. Kärnan i MAPIR-projektet är att utforska sätt att hantera dessa utmaningar i det fornsvenska materialet.

Finansierat av Marcus och Amalia Wallenbergs stiftelse

nya projekt: distributionella metoder

Forskning

Infrastruktur	>
SweFN++	>
META-NORD	>
KELLY	>
Kulturomik	>
CONPLISIT	>
Digital areallingsvistik	>
ITG	>
MOLTO	>
PINCORE	>
A System Architecture for ICALL	>
Akademiska ordlistor	>
Corpus-driven induction of linguistic knowledge	>

Corpus-driven induction of linguistic knowledge

The project aims to find automatic, corpus-based methods for inducing linguistic constructions and semantic frames, and representing their meaning using distributional semantics. In addition, the project will study the interaction between the automatically induced meaning representations and symbolic, knowledge-based resources such as frame and construction inventories, and use the representations in natural language processing (NLP) tools. It will combine two recent developments in unsupervised NLP: distributional methods for building and processing geometric meaning representations from corpora, and unsupervised semantic frame and role induction.

The results of the project will advance research in NLP and have practical benefits in applications: Corpus-induced semantic representations will be able to move beyond single words, and be formalized in terms of frame semantics and construction linguistics. Automatic syntactic and semantic analysis tools can be made more robust since they can use linguistic information beyond the word level. Linguistic resource building will benefit by the automatic methods for construction and frame discovery that the project will devise. NLP applications such as information extraction, opinion mining, grammar checking, and computer-assisted language learning can integrate the semantic frames and linguistic constructions discovered by the project, and use their distributional representations to understand their meaning.

The project is funded by the Swedish Research Council, grant 2013-4944, *Distributional Methods to Represent the Meaning of Frames and Constructions*, and lasts between 2014 and 2018.

Staff:

- Richard Johansson

Finansierat av VR

nya projekt: SWE-CLARIN

- CLARIN: ESFRI-förberedelsefas 2008-01 – 2011-06
- 9 svenska medlemmar (varav 2 partners)
- CLARIN ERIC startade 29/2 2012 med 9 medlemmar
- SWE-CLARIN-ansökan beviljad av VR 2013.
- Mål för SWE-CLARIN:

1. bilda en svensk nod i CLARIN ERIC:

- Göteborgs universitet (Språkbanken, SND)
- KTH (TMH)
- Linköpings universitet (NLP-lab)
- Lunds universitet (Humanistlaboratoriet)
- Stockholms universitet (datorlingvistik)
- Uppsala universitet (datorlingvistik)
- Språkrådet
- DigiSam

2. bygga en basinfrastruktur för CLARIN i Sverige

CLARIN-conceptet

- e-vetenskap – i form av språkteknologi som forskningsverktyg – för discipliner där text (och tal) är primärdata:
 - humaniora
 - samhällsvetenskap
 - (vissa sorters) medicin
- CLARINs betydelse växer i takt med digitaliseringen av kulturarvet och den elektroniska kommunikationens utbredning

digital spetsforskningspotential

Ökat intresse för gamla gruvor

Publicerat: måndag 02 juli 2007 kl 10:22. Nyheter P4 Norrbotten | 3 Dela ▼



Prospektering.

Ny och effektivare teknik har gjort att intresset för gamla nedlagda gruvor har ökat markant. Lavergruvan inom Älvsbyns kommun är ett sådant exempel. Hos Bergstaten som handlägger prospekterings- och gruvfrågor ser man en stor anhopning av undersökningstillstånd i anslutning till gamla fyndigheter.

Precis som vid gruvbrytning, kräver stora mängder 'informationsglast' digitalt text- och talmaterial effektiv teknik för sökning, korrelering och korsindexering i det språkliga innehållet – även mellan språk – för att forskningen ska få ut användbara primärdata ur det.

Men bara som man kan fråga får man svar, så planerna för Språkbanken handlar om att kunna erbjuda nya sorters svar:

- ▶ korpusjämförelser
- ▶ namntagging
- ▶ textmetadata
- ▶ syntaktisk sökning
- ▶ sökvisualisering (t.ex. trender, kartor)
- ▶ smartare träffgruppering, t.ex. visningssortering efter 'semantisk' kontext
- ▶ bättre syntaxanalys
- ▶ annotering av historiska material
- ▶ talspråk och ljud
- ▶ även annan forskning än språkvetenskap
- ▶ korpusvarieteter (användningar och användare), men även andra gränssnitt (med gemensamma nättjänster i bakändan)

Vi också gärna veta vilka frågor forskare och andra vill kunna ställa till materialet.

tack!

